Inférence semi-automatique et interactive de règles sans vérité terrain

Cérès Carton, Aurélie Lemaitre, Bertrand B. Coüasnon

Inférence semi-automatique et interactive de règles sans vérité terrain[1]

Cérès Carton* — **Aurélie Lemaitre**** — **Bertrand Coüasnon***

* *IRISA - INSA*
Université Européenne de Bretagne
Campus de Beaulieu
35042 Rennes Cedex, France
couasnon@irisa.fr
** *IRISA - Université Rennes 2*
Université Européenne de Bretagne
Campus de Beaulieu
35042 Rennes Cedex, France
aurelie.lemaitre@irisa.fr

RÉSUMÉ. *La conception de systèmes de reconnaissance de documents à partir de documents non annotés est particulièrement difficile. En général, les méthodes statistiques ne peuvent apprendre sans une vérité terrain annotée, contrairement aux méthodes syntaxiques. Cependant, pour ces dernières, leur capacité à ne pas nécessiter de données annotées est due au fait que la description du document est réalisée manuellement par le concepteur. L'adaptation à un nouveau type de documents est alors fastidieuse car l'ensemble du processus manuel d'extraction de connaissance doit être refait. Dans cet article, nous proposons une méthode pour extraire de la connaissance et générer des règles sans aucune vérité terrain. En utilisant de grands volumes de documents non annotés, il est possible d'étudier les redondances existantes sur des éléments extraits des images de documents. La redondance est exploitée grâce à un clustering automatique. Une interaction utilisateur permet d'apporter des informations sémantiques aux clusters ainsi détectés. Dans les travaux présentés, les éléments extraits sont des mots-clés détectés à l'aide de word spotting. Cette approche a été appliquée à la localisation de champs dans des registres de mariages anciens, issus de la base de documents de la compétition HIP2013 FamilySearch. Les résultats obtenus montrent que nous avons pu automatiquement inférer des règles à partir de documents non annotés, en exploitant la redondance d'éléments extraits de ces documents.*

1. Traduction étendue d'un article publié en anglais à ICDAR 2015 (Carton *et al.*, 2015)

ABSTRACT. *Dealing with non annotated documents for the design of a document recognition system is not an easy task. In general, statistical methods cannot learn without an annotated ground truth, unlike syntactical methods. However their ability to deal with non annotated data comes from the fact that the description is manually made by a user. The adaptation to a new kind of document is then tedious as the whole manual process of extraction of knowledge has to be redone. In this paper, we propose a method to extract knowledge and generate rules without any ground truth. Using large volume of non annotated documents, it is possible to study redundancies of some extracted elements in the document images. The redundancy is exploited through an automatic clustering algorithm. An interaction with the user brings semantic to the detected clusters. In this work, the extracted elements are some keywords extracted with word spotting. This approach has been applied to old marriage record field detection on the FamilySearch HIP2013 competition database. The results demonstrate that we successfully automatically infer rules from non annotated documents using the redundancy of extracted elements of the documents.*

MOTS-CLÉS : *Reconnaissance de documents structurés, Inférence de règles, Extraction de connaissances, Partitionnement de données, Données non annotées.*

KEYWORDS: *Document structure recognition, Rule inference, Knowledge extraction, Clustering, Non annotated data.*

1. Introduction

L'apprentissage automatique d'un système de reconnaissance de structures de documents et la capacité à gérer le manque de bases de documents annotés sont deux problèmes particulièrement difficiles lorsqu'ils sont pris séparément. La difficulté est encore plus grande lorsqu'il s'agit de gérer ces deux problèmes conjointement. Or, à notre connaissance, il n'existe pas de méthode capable de le faire dans la littérature.

Apprendre automatiquement un système de reconnaissance de structures de documents est, en général, le domaine privilégié des méthodes statistiques. Cette étape d'apprentissage les rend facilement adaptables à la reconnaissance d'un nouveau type de documents. Cependant, ces méthodes ont besoin d'une vérité terrain annotée pour l'étape d'apprentissage. C'est le cas pour les méthodes à base de CRF (Conditional Random Fields)(Montreuil *et al.*, 2010), de champs bi-dimensionnels markoviens (Lemaitre *et al.*, 2007) ou basées sur l'algorithme Espérance-Maximisation (Cruz et Terrades, 2014). Malheureusement, une vérité terrain n'est pas toujours disponible pour apprendre un nouveau système de reconnaissance de structures de documents. En effet, la génération des données de vérité terrain est une tâche laborieuse et coûteuse, car elle implique une forte intervention humaine. Des méthodes ont été proposées pour synthétiser des données annotées en applicant des modèles de dégradations (Kieu *et al.*, 2013), mais ces modèles ne sont pas adaptés pour la reconnaissance de la structure du document et ne peuvent pas produire de structures réalistes. À ce jour, il n'y a pas de solution dans la littérature permettant de se passer d'une vérité terrain annotée pour les méthodes de reconnaissance statistiques.

À l'inverse, les méthodes syntaxiques sont en général en mesure de faire face à des bases de documents non annotées. Cependant, cette capacité provient du fait que les connaissances sur les documents ne sont pas apprises automatiquement, mais manuellement et explicitement exprimées sous forme de règles par l'utilisateur (Coüasnon, 2006). Ainsi, ces méthodes ne peuvent pas être facilement adaptées à un nouveau type de documents puisque l'ensemble de l'extraction manuelle de la connaissance doit être refait. Des méthodes d'inférence de règles existent et ont été étudiées dans de nombreux domaines (de la Higuera, 2005), mais principalement mono-dimensionnels car c'est une tâche particulièrement complexe dans un contexte bi-dimensionnel. Shilman (Shilman *et al.*, 2005) présente une méthode pour apprendre des modèles grammaticaux non-génératifs pour l'analyse du document, en se concentrant sur la sélection de caractéristiques et l'estimation de paramètres. Cependant, cette méthode a toujours besoin d'une vérité terrain annotée pour définir tous les paramètres du modèle.

Dans le domaine de l'analyse de documents, des méthodes de classification de documents ont été développées ne nécessitant pas de vérité terrain ou uniquement très peu d'exemples annotés. Dans l'objectif de diminuer le nombre de document annotés nécessaires, Rusiñol (Rusinol *et al.*, 2013) propose une méthode incrémentale qui ne nécessite qu'une seule image annotée par classe. Cette méthode s'appuie sur un lo-

giciel d'OCR. Elle fonctionne donc bien si les documents ne sont pas trop dégradés, afin de permettre à l'OCR d'obtenir de bons taux de reconnaissance, et si l'ensemble des classes de documents à classer est connu par avance. Cependant, les méthodes de classification automatique de documents ne permettent pas d'atteindre notre objectif, puisqu'elles n'ont pas été conçues pour générer des règles décrivant la structure des documents.

Dans cet article, nous proposons une méthode pour extraire des connaissances et en déduire des règles à partir de documents non annotés. L'analyse est basée sur l'étude des redondances d'éléments extraits des documents dans de grandes bases de documents. Le principal avantage d'une telle méthode est de ne pas nécessiter de vérité terrain, tout en assurant une extraction semi-automatique et interactive de la connaissance. Dans la section 2, nous présentons un aperçu de la méthode proposée. Afin de montrer l'efficacité de notre approche lorsqu'il n'y a pas de vérité terrain connue sur les documents, nous proposons d'appliquer notre méthode à l'analyse de registres de mariages mexicains. Les données sont issues du corpus de la compétition HIP2013 FamilySearch.

Nous présentons dans la section 3, le corpus HIP2013 FamilySearch et nous détaillons la tâche que nous nous sommes fixée, qui diffère de la tâche du concours. Nous présentons ensuite en section 4, la phase de création de la pseudo vérité terrain ainsi que dans la section 5, l'inférence des modèles et la construction de la description grammaticale. Les résultats obtenus avec cette description grammaticale inférée sans vérité terrain, sont évalués dans la section 6, par rapport à ceux obtenus avec la description grammaticale définie manuellement et soumise au concours en 2013. Enfin, nous concluons sur les apports de notre méthode sur ce corpus.

2. Présentation générale de la méthode

Afin de permettre une extraction automatique de connaissances à partir de documents non annotés, nous proposons d'étudier les redondances d'éléments extraits dans les documents dans un large volume de documents. Les éléments extraits, que nous nommons également primitives, sont produits en utilisant un système de reconnaissance qui ne se base pas sur une connaissance de la structure des documents. Ces éléments peuvent être variés : résultats issus d'un OCR, mots-clés détectés à l'aide de *word spotting*, lignes ou blocs de texte, segments de lignes, etc. Ces éléments extraits correspondent aux terminaux de la grammaire dont nous voulons inférer les règles.

L'une des caractéristiques majeures de ces éléments extraits est qu'ils ne sont pas totalement fiables. Parmi les éléments extraits, certains correspondent effectivement à ce que l'utilisateur recherche tandis que d'autres ne correspondent pas. En conséquent, l'extraction de connaissances ne peut pas être effectuée directement sur les éléments extraits. Nous devons d'abord procéder à une phase de *fiabilisation des données* afin de construire une pseudo vérité terrain. Ce sont ces éléments fiabilisés qui sont ensuite utilisés pour l'extraction des connaissances et l'inférence des règles. La fiabilisation

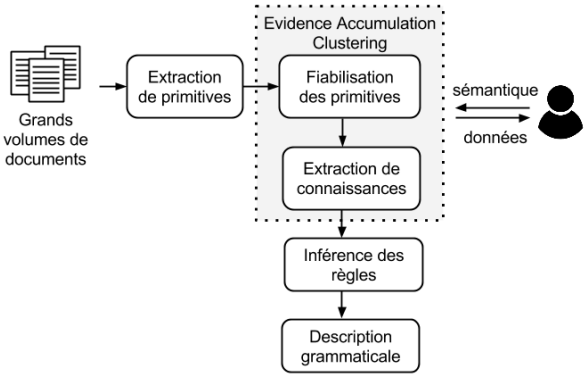


Figure 1. *Schéma général de la méthode d'inférence de règles sans vérité terrain*

des données et l'extraction des connaissances sont toutes deux basées sur la combinaison de méthodes de clustering, qui font émerger automatiquement des structures, et d'une interaction avec l'utilisateur, qui apporte du sens aux structures détectées automatiquement. Nous détaillons ces étapes dans la suite de cet article.

3. Présentation des données

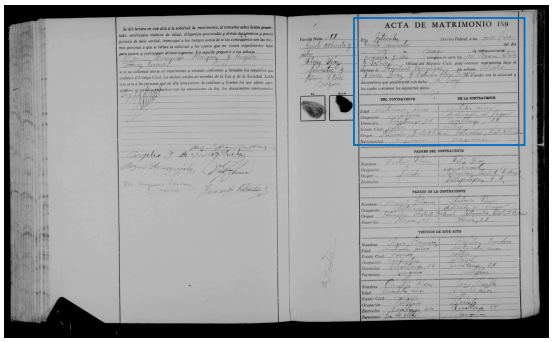
3.1. *Concours HIP2013 FamilySearch*

La tâche du concours HIP2013 FamilySearch consiste à détecter quatre régions d'intérêt dans des actes de mariages mexicains d'archives (XX^e siècle). Ces régions d'intérêt sont des champs manuscrits dans un texte pré-imprimé et correspondent aux informations suivantes (figure 2) :

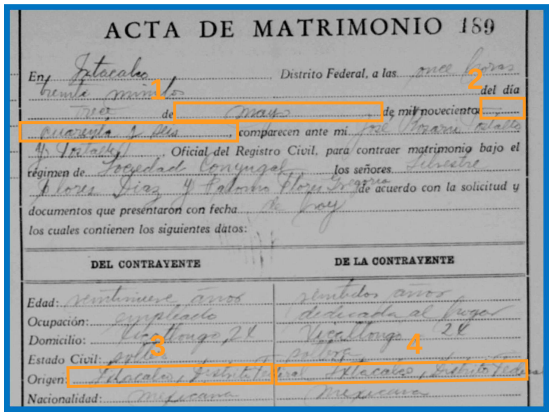
1) Mois de mariage
2) Année de mariage
3) Ville d'origine de l'époux
4) Ville d'origine de l'épouse

Après avoir détecté les régions d'intérêt, les participants doivent regrouper les images selon le contenu textuel des régions d'intérêt. Chaque image est regroupée quatre fois, une fois pour chacune des régions d'intérêt. Il n'est par contre pas nécessaire de reconnaitre le contenu manuscrit des régions d'intérêt.

Pour la compétition, un jeu de données d'apprentissage de 10 000 documents était fourni aux participants. Une vérité terrain était fournie pour chaque document sous la forme du contenu textuel des régions d'intérêt. La vérité terrain ne contient pas la localisation des régions d'intérêt. Le jeu de données de test est composé de 20 000 images.



(a) Exemple d'acte de mariage mexicain à analyser. La zone entourée en bleu est la zone contenant les 4 régions d'intérêt



(b) Représentation sur un exemple des quatres zones d'intérêt à détecter

Figure 2. *Exemple d'acte de mariage mexicain du corpus HIP2013 FamilySearch*

3.2. *Sous-tâche du concours*

Dans notre expérimentation, nous nous sommes focalisés sur une sous-tâche de la compétition : la localisation des régions d'intérêt *mois* et *année*. En effet, nous désirons évaluer notre méthode indépendamment des performances de la méthode de clustering sur l'écriture manuscrite qui vient en post-traitement du système de reconnaissance de structure de documents. Nous cherchons donc à localiser les zones d'intérêt, sans évaluer la reconnaissance de leur contenu. La zone est détectée qu'il y ait un écrit présent ou non dans le champ. Pour cette tâche, il n'y a pas de vérité terrain disponible sur le corpus du concours HIP2013 FamilySearch. L'apprentissage a été réalisé sur 7 000 documents du jeu de données d'apprentissage sans vérité terrain du concours.

4. Création de la pseudo vérité terrain

Les documents utilisés dans la base d'apprentissage ne possèdent pas de vérité terrain annotée manuellement. Pour permettre l'inférence automatique et interactive de la description grammaticale, nous devons construire une pseudo vérité terrain. Pour cela, nous procédons en deux étapes que nous allons détailler ci-après :

1) Extraction des primitives ;

2) Fiabilisation des primitives.

4.1. *Extraction des primitives*

Les primitives utilisées sont des mots-clés du paragraphe contenant les régions d'intérêt *mois* et *année*. Nous utilisons huit mots-clés différents (figure 3) : *Distrito* (district), *Federal* (fédéral), *día* (jour), *de* (de), *de mil novecientos* (de mille neuf cents), *comparecen* (comparaissent), *Oficial* (officier) et *Registro* (bureau de l'état civil). Nous cherchons à obtenir pour chaque document une unique occurrence de chaque mot-clé. Les mots-clés ont été sélectionnés car ils sont présents dans le paragraphe contenant les champs *mois* et *année* que nous cherchons à localiser. Ces mots-clés nous permettront de délimiter la position des deux champs recherchés.

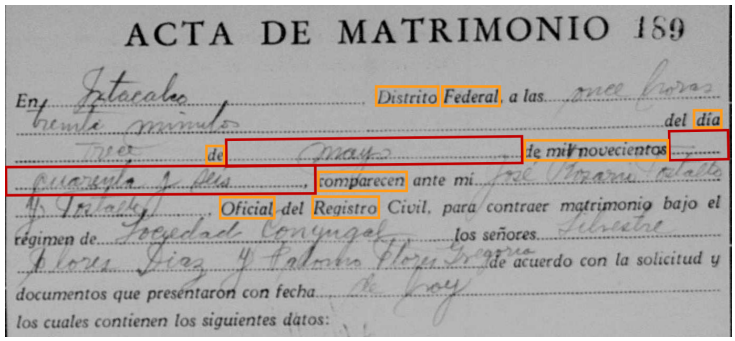


Figure 3. *Mots-clés utilisés comme primitives d'analyse pour la construction de la pseudo vérité terrain*

Les mots-clés sont détectés selon une méthode basée sur la disposition de descripteurs locaux appelés Points d'Intérêt (POI) (Camillerapp, 2012). Nous représentons ici succinctement les grands principes de cette approche. Nous n'avons pas utilisé d'OCR car la dégradation des documents ainsi que les interactions manuscrits/imprimés ne permettaient pas une reconnaissance suffisante des mots-clés.

Les POI sont des points de l'image qui présentent des variations locales de luminosité. Cette sélection doit être stable : les mêmes points sont sélectionnés dans toutes les images pour représenter le même objet. De plus, ces points doivent être discriminants : dans une image, il doit y avoir peu de confusion entre les descripteurs locaux. Pour

chaque point d'intérêt sélectionné, nous calculons un descripteur local. Nous utilisons le descripteur introduit par Lowe (Lowe, 2004). Ce descripteur calcule des statistiques sur la direction du gradient dans un petit voisinage du point.

Mot-clé	Nombre d'occurrences par page
comparecen	4,0
de	9,1
de mil novecientos	1,1
día	6,5
Distrito	2,7
Federal	2,0
Oficial	3,7
Registro	1,0

Tableau 1. *Nombre moyen d'occurrences des mots-clés par document détectés par la méthode des points d'intérêt (POI) sur la base d'apprentissage de 7 000 documents*

Les mots-clés obtenus lors de l'extraction des primitives sont bruités. Pour un document, nous n'obtenons pas une unique occurrence de chaque mot-clé (tableau 1). Par exemple pour les mots-clés courts « día » et « de », nous trouvons en moyenne respectivement 6,5 et 9,1 occurrences par document. Afin d'obtenir une pseudo vérité terrain pour l'inférence des règles grammaticales, il est indispensable de fiabiliser les primitives, c'est-à-dire de sélectionner les mots-clés détectés qui correspondent à la présence réelle du texte cherché.

Pour cela, nous utilisons un algorithme de clustering : le clustering permet de mettre en avant des clusters d'éléments similaires et l'utilisateur vient apporter du sens à chacun de ces clusters. Le choix de l'algorithme de clustering est déterminé par plusieurs contraintes. Tout d'abord, l'algorithme choisi doit être capable de s'adapter à de nombreux jeux de données différents afin de garantir la généricité de la méthode. Ensuite, nous n'avons pas d'*a priori* sur les données, l'algorithme choisi doit donc avoir le moins possible de paramètres à déterminer manuellement, et en particulier nous ne devons pas avoir à spécifier le nombre de clusters. Enfin, nous utilisons cet algorithme en deux points distincts de l'analyse : pour la fiabilisation des primitives et pour l'extraction de connaissance (figure 1). Pour cela, il est intéressant que l'algorithme détermine automatiquement le nombre de clusters optimal mais également de pouvoir facilement sur-segmenter les données, afin de faciliter la fiabilisation des primitives.

4.2. *Evidence Accumulation Clustering*

En tenant compte de ces contraintes, nous avons sélectionné la méthode de clustering par accumulation de preuves (Evidence Accumulation Clustering) introduite par Fred et Jain (Fred et Jain, 2002). L'EAC clustering construit tout d'abord N partitions

avec différents algorithmes de clustering. Dans notre implémentation, nous avons utilisé l'algorithme des K-moyennes avec une valeur aléatoire pour le paramètre K. Ces différentes partitions sont combinés pour générer une matrice de similarité $\mathcal{M}$:

$$\mathcal{M}(i,j) = \frac{n_{ij}}{N}$$

où n_{ij} correspond au nombre de fois où les éléments i et j sont dans le même cluster. La partition finale est obtenue à l'aide d'un algorithme de clustering hiérarchique pour lequel le critère du temps de vie maximum (*maximum cluster lifetime criterion*) est utilisé. Nous découpons le dendrogramme au niveau de la partition qui perdure le plus longtemps. Il est alors facilement possible de sur-segmenter les données en choisissant un autre point de découpe du dendrogramme.

4.3. *Fiabilisation des primitives*

Comme nous l'avons indiqué dans la section 4.1, les primtives obtenues ne sont pas fiables. Nous présentons ici la méthode employée pour améliorer leur fiabilité.

4.3.1. *Clustering*

Le clustering est effectué directement sur les éléments extraits des images de documents. Ces éléments peuvent être très variés et dépendent de l'objectif de la construction de la pseudo vérité terrain. Par exemple, des propriétés de taille, de position, d'espacements sur des mots, des segments de droites, des lignes de textes, etc. Ils peuvent être extraits par un extracteur de primitives, un OCR, du word spotting, un système de reconnaissance de structure de documents existant ou en cours de construction, etc. L'algorithme de clustering (cf. section 4.2) produit une partition des données qui est utilisée afin de supprimer rapidement et efficacement des clusters entiers de données ne contenant pas des occurrences d'intérêt pour l'étape d'extraction de connaissances.

4.3.2. *Interaction utilisateur*

Lorsque la partition a été automatiquement construite sur les données, chaque cluster est visualisé par l'utilisateur. C'est une étape essentielle de la construction de la pseudo vérité terrain car l'utilisateur apporte ici du sens aux données, en séparant les éléments qu'il considère comme du bruit, des éléments à conserver pour un apprentissage. Pour cela, il visualise chaque cluster et détermine si le cluster est conservé ou non dans l'analyse. Afin d'effectuer cette étape efficacement, seuls cinq exemples représentatifs du cluster sont présentés à l'utilisateur. Les exemples considérés comme représentatifs sont ceux situés près du centroïde du cluster. Une mesure de la variabilité intra-cluster est également donnée à l'utilisateur afin d'assurer la représentativité des exemples présentés. Plus la variabilité intra-cluster est faible, plus l'utilisateur peut avoir confiance dans sa décision.

Dans l'exemple des actes de mariage mexicains, la fiabilisation des primitives est effectuée en construisant les clusters à partir des positions normalisées des mots clés

dans la zone de l'acte de mariage définie par le paragraphe accompagné du titre « Acta de matrimonio ». Chaque cluster est ensuite conservé ou non après une interaction utilisateur en visualisant quelques exemples représentatifs du cluster. La fiabilisation se fait indépendamment pour chacun des huit types de mots-clés. À la fin de cette étape de fiabilisation des primitives, les données sont utilisées pour l'extraction des connaissances.

5. Extraction de connaissance : construction de la description grammaticale

Une fois la pseudo vérité terrain constituée, nous pouvons procéder à l'extraction de connaissance afin de construire la description grammaticale permettant de décrire les registres de mariages mexicains. Nous cherchons ici à apprendre les différents types de pré-imprimés existant dans le corpus pour les registres de mariage. Lorsqu'un registre est analysé, nous pouvons alors chercher dans cet ensemble de modèles pré-imprimés connus celui qui a été utilisé pour constituer le document.

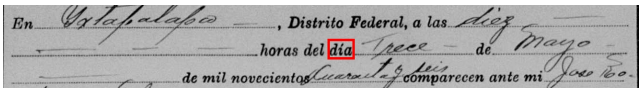
5.1. *Inférence des modèles de mots-clés*

Un clustering des positions des mots-clés dans la page selon l'axe des abscisses et l'axe des ordonnées sur la base d'apprentissage de 7 000 documents est effectué sur les données de la pseudo vérité terrain. Le clustering nous permet de détecter les modèles de position par mot-clé.

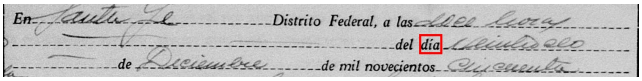
Mot-clé	Nombre de modèles de position
comparecen	9
de	5
de mil novecientos	6
día	7
Distrito	5
Federal	4
Oficial	7
Registro	4

Tableau 2. *Nombre de modèles de position détectés pour chaque mot-clé durant la phase d'extraction de connaissance sur la base d'apprentissage de 7 000 documents*

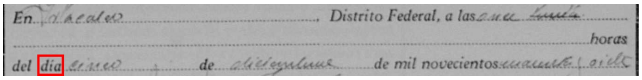
Les clusters ainsi formés sont visualisés par l'utilisateur qui apporte du sens dans l'analyse en proposant des libellés de cluster porteurs d'une sémantique. Il valide ainsi la pertinence du clustering produit automatiquement. Le nombre de modèles de position pour chaque mot-clé est présenté dans le tableau 2. La figure présente trois exemples de modèles de position différents détectés pour le mot-clé « día ».



(a) Modèle 1



(b) Modèle 2



(c) Modèle 3

Figure 4. *Exemples de trois modèles de position différents pour le mot-clé « día » détectés dans le corpus d'apprentissage HIP2013 FamilySearch*

5.2. *Inférence des modèles de documents*

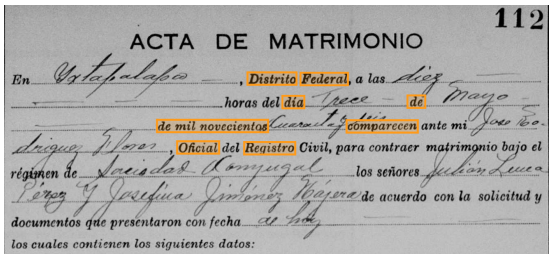
Nous construisons pour chaque document de l'ensemble d'apprentissage ainsi constitué une signature du document. Nous utilisons pour cela les modèles de position par mot-clé. L'occurrence du mot-clé contenu dans le document est affecté au modèle de position correspondant et la signature est la concaténation de chacun des modèles de position correspondants. Un exemple est présenté dans le tableau 3.

	Modèle de position du mot-clé				
Fichier	dia	de	de mil nov	comparecen	signature
00001	3	4	7	6	3#4#7#6

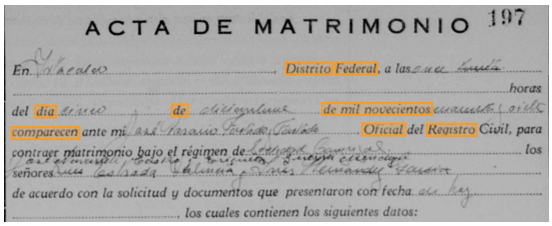
Tableau 3. *Exemple de création de la signature pour un registre de mariage*

Pour construire les modèles de documents, nous utilisons les registres pour lesquels une unique occurrence de chacun des huit mots-clés est présente. Ce sont les documents pour lesquels il n'existe pas d'ambiguïté sur les modèles de position par mot-clé à utiliser. L'apprentissage des modèles de documents a pu être effectuée sur 5406 documents sur les 7000 documents de l'ensemble d'apprentissage.

Une analyse des fréquences des signatures des documents est ensuite effectuée. 11 modèles de pré-imprimés différents sont alors détectés dans l'ensemble d'apprentissage, alors qu'aucune information n'est disponible *a priori* sur le nombre de pré-imprimés existants dans la base. La figure 5 présente deux exemples de modèles de pré-imprimés différents détectés dans le corpus d'apprentissage. Ces modèles sont construits sans être perturbés par des erreurs sur la détection des mot-clés, puisque les mots-clés utilisés pour l'apprentissage ont été fiabilisés préalablement (les mots-clés doivent appartenir à un cluster et donc se retrouver dans des positions similaires sur un certain nombre de document, et le cluster a été validé par l'utilisateur).



(a) Modèle 1



(b) Modèle 2

Figure 5. *Exemples de deux modèles de pré-imprimés différents détectés dans le corpus d'apprentissage HIP2013 FamilySearch*

La répartition des 11 modèles de documents dans le corpus d'apprentissage est présenté dans le tableau 4. Leur répartition inégale montre l'intérêt d'une analyse automatique et exhaustive de l'ensemble d'apprentissage. En effet, il aurait été très difficile avec une analyse manuelle de détecter l'ensemble de ces modèles. L'analyse aurait alors été effectuée sur un petit ensemble d'apprentissage qui n'aurait pas pu être représentatif.

Model	1	2	3	4	5	6	7	8	9	10	11
Count	1448	822	740	652	566	470	359	123	92	33	25

Tableau 4. *Les modèles de documents sont inégalement répartis dans le corpus d'apprentissage*

5.3. *Intégration dans la description grammaticale*

Les modèles de pré-imprimés détectés et validés en interaction avec l'utilisateur nous permettent de générer automatiquement la description grammaticale des registres de mariage. Pour cela, nous avons besoin de générer les opérateurs de position de chacun des mots-clés par modèle de pré-imprimé. Un opérateur de position défini un point de vue (2 coordonnées) et une zone de recherche (elle-même définie par 2 points, soit 4 coordonnées). L'opérateur de position permet au mécanisme d'analyse grammatical

de sélectionner des éléments appartenant à la zone de recherche selon un ordre donné par la distance euclidienne de l'élément au point de vue. Nous inférons donc automatiquement : 6 paramètres (pour chaque opérateur de position) $\times$ 8 opérateurs de position (un par position de mot-clé dans le formulaire) $\times$ 11 modèles de pré-imprimés = 528 paramètres.

Afin de déterminer pour un document à analyser quel modèle de pré-imprimé a été analysé, nous utilisons l'opérateur FIND_BEST_FIRST introduit par Maroneze (Maroneze *et al.*, 2011). Cet opérateur permet de construire une grammaire stochastique localement. Lorsque nous analysons un document, chaque modèle de pré-imprimé est testé. Une pénalité est alors calculé représentant la non-adéquation du modèle de pré-imprimé au document analysé. L'opérateur FIND_BEST_FIRST nous permet alors de sélectionner le modèle de pré-imprimé avec la pénalité la plus faible, c'est-à-dire celui qui correspond le mieux au document.

Calcul de la pénalité Lorsque nous testons l'adéquation d'un modèle, chacun des mots-clés est recherché à sa position supposée, apprise sur l'ensemble d'apprentissage. Si le mot-clé n'est pas trouvé à cette position, la pénalité du modèle est augmentée de 1. Si le mot-clé est trouvé alors la pénalité du modèle est augmenté de :

$$1 - \frac{\text{aire de l'intersection}}{\text{aire du mot-clé}}$$

La figure 6 montre un exemple de calcul de pénalité pour un mot-clé. Le mot-clé « de mil novecientos » est recherché dans la zone rouge. Une occurrence est trouvée qui n'est pas totalement incluse dans la zone de recherche. La pénalité calculée pour ce mot-clé est 0,477444.

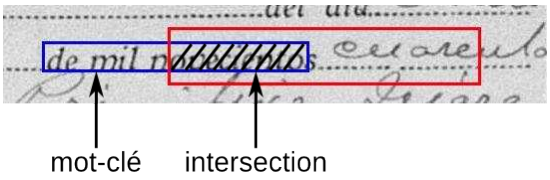


Figure 6. *La pénalité du mot clé « de mil novecientos » est 0,477444 car il n'est pas totalement inclus dans la zone de recherche*

Cette approche nous permet à la fois de détecter les bons mots-clés parmi les mots-clés contenant du bruit détectés avec les modèles de POI. De plus, cela nous permet également de synthétiser le mot-clé s'il n'y a pas de primitives correspondant dans le document.

6. Évaluation expérimentale

L'évaluation est faite sur 2 000 documents du jeu de données de test du concours qui ont été annotés manuellement. Dans chaque document, la position des champs

« mois » et « année » sont annotés. Nous ne tenons pas compte pour cela de la présence ou non de texte dans les champs.

6.1. *Métrique*

Pour évaluer la sous-tâche de détection de la position des champs « mois » et « année », nous devons évaluer la correspondance spatiale entre les champs détectés avec notre système de reconnaissance et ceux annotés dans la vérité terrain. Pour cela, nous utilisons la métrique introduite par Garris (Garris, 1995) qui nous permet d'évaluer le recouvrement entre la zone attendue et la zone obtenue. Nous calculons l'intersection entre la zone attendue et la zone effectivement obtenue. La largeur de cette intersection doit être proche de la largeur de la zone attendue. La hauteur de l'intersection doit être également suffisamment grande pour pouvoir contenir le texte du champ.

Nous définissons deux seuils pour l'évaluation des résultats :

– un champ est considéré comme *complètement reconnu* si au moins 95% de la largeur et 75% de la hauteur a été reconnu

– un champ est considéré comme *partiellement reconnu* si 1) il n'est pas *complètement reconnu*, et 2) au moins 80% de sa largeur est reconnu ainsi que 75% de sa hauteur.

Dans les autres cas, le champ est considéré comme manquant.

Le document est considéré comme reconnu lorsque tous les champs du document sont complètement ou partiellement reconnus et qu'il n'y a pas de zone détectée en trop dans le document.

6.2. *Résultats*

Lors de l'évaluation, nous comparons les résultats obtenus par notre méthode à la vérité terrain. Les résultats présentés dans le tableau 5 montrent que la description grammaticale construite à partir des modèles de pré-imprimés inférés permet de localiser efficacement les champs « mois » et « année ». Seulement 2,4% des champs ne sont pas reconnus et 89,8% des documents sont correctement reconnus. Cela montre que la description grammaticale inférée sans vérité terrain est efficace.

Nous comparons également les résultats obtenus avec ceux obtenus par une méthode où les modèles de documents sont définis manuellement. Cette méthode est celle soumise par Lemaitre (Lemaitre et Camillerapp, 2013) lors de la compétition HIP2013 FamilySearch. Dans cette description, quatre modèles ont été manuellement décrits. Cette méthode a été classée deuxième lors de la compétition HIP2013 FamilySearch.

Lorsque nous comparons les résultats obtenus par les modèles inférés aux modèles définis manuellement, nous pouvons remarquer que notre méthode obtient de meilleurs résultats. En effet, il y a moins de zones manquantes avec notre méthode

		Modèles	
		Inférés	Manuels
Zone	Reconnaissance complète	91.4%	89.7%
	Reconnaissance partielle	6.2%	4.0%
	Manquant	**2.4%**	6.3%
Taux de reconnaissance du document		**89.8%**	78.9%

Tableau 5. *Comparaison des résultats obtenus sur 2 000 documents avec des modèles inférés automatiquement et des modèles définis manuellement*

(131 zones manquantes contre 343 zones avec les modèles définis manuellement). De plus, le taux de documents bien reconnus est fortement amélioré en utilisant les modèles de pré-imprimés inférés, avec une augmentation de 11% de documents bien reconnus (soit 217 documents sur 2 000).

7. Conclusion

Dans cet article, nous avons présenté une méthode pour inférer des règles sur la structure de documents, à partir de documents non annotés. Pour pallier le manque de vérité terrain annotée, nous avons proposé de nous appuyer sur l'analyse des redondances d'éléments extraits dans un grand nombre de documents. Cette analyse combine une fiabilisation des éléments extraits pour construire une pseudo vérité terrain, avec une extraction automatique de connaissances. Pour chacune de ces deux tâches, nous avons utilisé avec succès la méthode de clustering par accumulation de preuves (Evidence Accumulation Clustering) introduite par Fred et Jain (Fred et Jain, 2002). Une interaction avec l'utilisateur permet en outre de donner une sémantique aux éléments produits par l'algorithme de clustering.

Nous avons validé notre approche sur la base de données de la compétition HIP2013 FamilySearch, en nous concentrant sur une sous-tâche de la compétition, la localisation des champs "mois" et "année". 2 000 documents de la base de données de test ont été annotés manuellement à cet effet et seront disponibles publiquement. Nos résultats montrent que nous pouvons efficacement extraire des connaissances à partir de documents non annotés, ce qui constitue une amélioration importante car l'annotation manuelle des documents est une étape très coûteuse dans la conception de systèmes de reconnaissance de documents. L'introduction d'un mécanisme itératif pourrait améliorer cette approche en détectant automatiquement les occurrences intéressantes qui auraient pu être supprimées au cours de la fiabilisation des données. Cette étape itérative pourrait également être utilisée pour détecter une nouvelle configuration de documents afin d'améliorer le système de reconnaissance de documents.

8. Bibliographie

Camillerapp J., « Utilisation des points d'intérêt pour rechercher des mots imprimés ou manuscrits dans des documents anciens », *Conférence Internationale sur l'Écrit et le Document (CIFED'12)*, p. 163-178, 2012.

Carton C., Lemaitre A., Coüasnon B., « Automatic and interactive rule inference without ground truth », *International Conference on Document Analysis and Recognition (ICDAR)*, Nancy, France, August, 2015.

Coüasnon B., « DMOS, a Generic Document Recognition Method : Application to Table Structure Analysis in a General and in a Specific Way », *International Journal on Document Analysis and Recognition, IJDAR*, vol. 8, n° 2, p. 111-122, June, 2006.

Cruz F., Terrades O. R., « EM-Based Layout Analysis Method for Structured Documents », *22nd International Conference on Pattern Recognition*, p. 315-320, 2014.

de la Higuera C., « A Bibliographical Study of Grammatical Inference », *Pattern Recogn.*, vol. 38, n° 9, p. 1332-1348, September, 2005.

Fred A. L. N., Jain A., « Data clustering using evidence accumulation », *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, p. 276-280 vol.4, 2002.

Garris M., « Evaluating spatial correspondence of zones in document recognition systems », *Image Processing, 1995. Proceedings., International Conference on*, vol. 3, p. 304-307 vol.3, Oct, 1995.

Kieu V. C., Journet N., Visani M., Domenger J.-P., Mullot R., « Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes », *International Conference on Document Analysis and Recognition (ICDAR 2013)*, Washington, DC, United States, p. 489-493, August, 2013.

Lemaitre A., Camillerapp J., « HIP 2013 FamilySearch Competition - Contribution of IRISA », *HIP - ICDAR Historical Image Processing Workshop*, Washington, United States, August, 2013.

Lemaitre M., Grosicki E., Geoffrois E., Preteux F., « Preliminary experiments in layout analysis of handwritten letters based on textural and spatial information and a 2D Markovian approach », *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, p. 1023-1027, Sept, 2007.

Lowe D. G., « Distinctive Image Features from Scale-Invariant Keypoints », *Int. J. Comput. Vision*, vol. 60, n° 2, p. 91-110, November, 2004.

Maroneze A. O., Coüasnon B., Lemaitre A., « Introduction of statistical information in a syntactic analyzer for document image recognition », *DRR*, p. 1-10, 2011.

Montreuil F., Nicolas S., Grosicki E., Heutte L., « A New Hierarchical Handwritten Document Layout Extraction Based on Conditional Random Field Modeling », *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, p. 31-36, Nov, 2010.

Rusinol M., Benkhelfallah T., D'Andecy V., « Field Extraction from Administrative Documents by Incremental Structural Templates », *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, p. 1100-1104, Aug, 2013.

Shilman M., Liang P., Viola P., « Learning nongenerative grammatical models for document analysis », *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, p. 962-969 Vol. 2, Oct, 2005.